TDT-76 (Deep Learning) Final Evaluation

Clearly print your full name HERE:

Answer all questions to the right of the arrows provided. **You must copy your answers onto the answer sheet at the end of this document to get credit for them.** Each correct answer is worth the same (one) point, while incorrect answers receive zero points. There are 22 total points on the exam, but you only need 20 correct answers to get a *perfect* score.

1 Questions

1. Which of the following is the most obvious situation for using a combination of convolution and recurrence in a deep neural network?

   (a) Generating captions for 30-second video clips.

   (b) Generating classical music in the style of a particular composer.

   (c) Playing three-dimensional checkers.

   (d) Predicting patient responses to a drug based on their age, gender, weight and blood type.

   (e) Recommending books for patrons of an online bookstore.

   $\Longrightarrow$ ..... (#1)

2. A deep neural network is tested on 1000 MRI images; 900 *negative* samples show no irregularities while 100 *positive* samples show obvious signs of disease D. If the network classifies 800 of the images as positive (and 200 as negative), and it incorrectly classifies 10 positive cases as negative, what are the accuracy, recall and precision values of this test:

   Accuracy $\Longrightarrow$ ..... (#2)

   Recall $\Longrightarrow$ ..... (#3)

   Precision $\Longrightarrow$ ..... (#4)

3. In Deep Learning, regularization is:

   (a) often achieved by using dropout

   (b) aided by dense hidden-layer representations

   (c) an attempt to reduce testing error while keeping training error low

   (d) an attempt to enhance generalization

   (e) often achieved by adding L_1 or L_2 terms to the loss function

   (f) All of the above (A-E)

   (g) Exactly 4 of (A-E)

(h) Exactly 3 of (A-E)

(i) Exactly 2 of (A-E)

(j) None of (A-E)

$\Longrightarrow \ldots\ldots$ (#5)

4. The most significant difference between Alpha Go and Alpha Go Zero is:

   (a) Alpha Go Zero uses Deep Reinforcement Learning

   (b) Alpha Go plays winning Go, whereas Alpha Go Zero is mainly a promising proof-of-concept system.

   (c) Alpha Go initially trains on expert-player data, whereas Alpha Go Zero learns primarily from self-play.

   (d) Alpha Go uses Q-Learning, while Alpha Go Zero uses SARSA.

   (e) Alpha Go Zero is being sued by Coca Cola for copyright infringement.

   $\Longrightarrow \ldots\ldots$ (#6)

5. A standard neural network classifier contains a total of 100 trainable weights (and no bias terms) and has an output layer consisting of 4 nodes. The loss function is mean-squared error (MSE). During training, it processes a total 500 minibatches, each of size 20. Calculate:

   The total number of gradients (of loss with respect to individual weights) computed by backpropagation $\Longrightarrow \ldots\ldots$ (#7)

   The total number of weight updates performed by backpropagation $\Longrightarrow \ldots\ldots$ (#8)

6. Which of the following statements is true about the equation below?

$$Y(j,k) = (X * w)(j,k) = \sum_{\gamma=-c}^{c} \sum_{\delta=-d}^{d} X(j+\gamma, k+\delta) w(\gamma, \delta) \tag{1}$$

   (a) This is the essence of 2-dimensional convolution.

   (b) The terms c and d in the summations represent strides.

   (c) When $c \neq d$, zero-padding becomes necessary.

   (d) There are $(2c + 2d + 2)$ weights in the kernel.

   (e) Both A and D

   (f) A, C and D

   (g) Both A and C

   (h) All of the above (A-D)

   (i) None of the above (A-D)

   $\Longrightarrow \ldots\ldots$ (#9)

7. Which of the following concepts is most important in understanding the computations behind the backpropagation algorithm?

   (a) The Bellman equation

   (b) The Hessian matrix

   (c) Boltzmann's equation

   (d) The Jacobian matrix

   (e) Adversarial training

   (f) Adam optimization

$\Longrightarrow$ ..... (#10)

8. Which of the following has **NOT** been a significant factor in the exceptional success of Deep Learning in the past 15 years.

   (a) ReLU
   (b) Geoffrey Hinton
   (c) GPU
   (d) Major changes to the original backpropagation algorithm
   (e) Yoshua Bengio
   (f) 2, 3 or 4 of the above (A-E)
   (g) None of the above (A-E)

   $\Longrightarrow$ ..... (#11)

9. In a deep neural network, layer U (of size m) sends its outputs to layer V (of size n) across weights W. Assume u_i is the output of the ith neuron of layer U, v_j is the output of the jth neuron of layer V, and $w_{i,j}$ is the weight on the connection from u_i to v_j. Layer U uses a hyperbolic tangent (tanh) activation function, while layer V uses a sigmoid activation function. What is the proper expression for the following partial derivative:

$$\frac{\partial v_j}{\partial u_i} = ? \tag{2}$$

   (a) $\sum_{i=1}^{m} u_i w_{i,j} v_j (1 - v_j^2)$
   (b) $\sum_{i=1}^{m} u_i w_{i,j} v_j (1 - v_j)$
   (c) $u_i w_{i,j} v_j (1 - v_j)$
   (d) $(1 - u_i^2) w_{i,j} v_j (1 - v_j)$
   (e) $\sum_{i=1}^{m} (1 - u_i^2) w_{i,j} v_j (1 - v_j)$
   (f) $\sum_{i=1}^{m} w_{i,j} v_j (1 - v_j)$
   (g) $w_{i,j} v_j (1 - v_j)$

   $\Longrightarrow$ ..... (#12)

10. In the same network as described in the previous question, what is the proper expression for the partial derivative

$$\frac{\partial v_j}{\partial w_{i,j}} = ? \tag{3}$$

   (a) $\sum_{i=1}^{m} u_i w_{i,j} v_j (1 - v_j^2)$
   (b) $\sum_{i=1}^{m} u_i v_j (1 - v_j)$
   (c) $u_i v_j (1 - v_j)$
   (d) $(1 - u_i^2) v_j (1 - v_j)$
   (e) $\sum_{i=1}^{m} (1 - u_i^2) w_{i,j} v_j (1 - v_j)$
   (f) $u_i w_{i,j} v_j (1 - v_j)$
   (g) $u_i (1 - u_i^2) v_j (1 - v_j)$

   $\Longrightarrow$ ..... (#13)

11. Which of the following statements about the cross-entropy loss function are true.

   (a) It is minimized by the conditional maximum likelihood estimate.
   (b) It should never be applied to softmax'ed output values.
   (c) It should never be applied to one-hot target vectors.

(d) It makes sense to use it when the outputs and the targets each represent probability distributions.

(e) It only makes sense to use it when the Hessian of the weight matrix has positive eigenvalues.

(f) All of the above (A-E)

(g) 4 of the above (A-E)

(h) 3 of the above (A-E)

(i) 2 of the above (A-E)

(j) None of the above (A-E)

$\Longrightarrow$ ..... (#14)

12. Which of the following statements are true about the equation below (in which t = the timestep or iteration):

$$\Delta w_{ij}(t) = -\lambda \frac{\partial Loss}{\partial w_{ij}} + \alpha \Delta w_{ij}(t-1) \tag{4}$$

(a) α is often set to 1 - λ for best results.

(b) This represents the core of Q-Learning, although λ and α are often replaced by a Taylor series.

(c) The choice of λ rarely affects the behavior of the overall algorithm.

(d) In rough search landscapes, choosing $\alpha = 0$ is often wise.

(e) The Adam optimizer incorporates the essence of this equation but uses more than just a single timestep of history to update w_{ij}.

(f) All of the above (A-E)

(g) 4 of the above (A-E)

(h) 3 of the above (A-E)

(i) 2 of the above (A-E)

(j) None of the above (A-E)

$\Longrightarrow$ ..... (#15)

13. A very useful embedding for a small collection of English words is created via a Word2Vec Skip-Gram Network with 6 hidden nodes. Given 4 words in this collection that are all quite different from one another semantically, which of the following is the most likely embedding for these 4 words?

(a) 111110, 101111, 111011, 110111

(b) 100001, 111000, 011110, 000111

(c) 101000, 001100, 011000, 010100

(d) 111000, 011100, 011000, 110000

(e) All 4 options (A-D) are equally likely.

$\Longrightarrow$ ..... (#16)

14. Autoencoding by a neural network is:

(a) trivial to achieve when the hidden layer and input layer have the same size.

(b) a key element of the Restricted Boltzmann Machine (RBM).

(c) a relatively simple task, but a useful support for more challenging tasks.

(d) useful for finding compression codes of the input patterns.

(e) easily regularized to produce sparse embeddings.

(f) All of the above (A-E)

(g) Some (but not all) of the above (A-E)

(h) None of the above

$\Longrightarrow$ ..... (#17)

15. A layer of 8 neurons, X ($x_1 \dots x_8$) sends its output to layer Y ($y_1 \dots y_5$), which uses a sigmoid activation function. Y is then compared to a target vector, T ($t_1 \dots t_5$), and the loss function is:

$$L = \frac{1}{2}\sum_{i=1}^{5}(y_i - t_i)^2 \tag{5}$$

- On the most recent forward pass, X produces values [0.3, 0.8, 0.1, 0.1, 0.1, 0.4, 0.7, 0.2],
- while Y outputs [0.9, 0.5, 0.2, 0.3, 0.1],
- and T = [1, 0, 0, 0, 0].
- Let w be the weight on the connection from x_1 to y_1.
- The learning rate is 0.2.

Using standard backpropagation, calculate the value of Δw based on this single forward pass.

$\Longrightarrow$ ..... (#18)

16. One of the most important connections between Deep Learning (DL) and Reinforcement Learning (RL) is:

(a) Deep networks serve as function approximators, thus replacing RL value tables.

(b) Bellman updates and SARSA often improve the backpropagation algorithm.

(c) Both DL and RL are essential tools for AI applied to board games and video games.

(d) By adding DL to RL, tasks that normally have sparse reinforcement gain plentiful reinforcement via the outputs of the deep network.

(e) None of the above

$\Longrightarrow$ ..... (#19)

17. Which of the following can be significant causes of the *vanishing gradient problem* in Deep Learning?

(a) Repeated products of W, the weights connecting the hidden layer to itself, in a recurrent network.

(b) The second derivative being equal to zero in an ReLU activation function.

(c) Saturation of the sigmoid activation function.

(d) Many hidden layers in the same network.

(e) Using the CPU instead of a GPU.

(f) All of the above (A-E)

(g) 4 of the above (A-E)

(h) 3 of the above (A-E)

(i) 2 of the above (A-E)

(j) None of the above (A-E)

$\Longrightarrow$ ..... (#20)

18. An Elman Network has a single hidden layer, H, and weight matrix W that fully connects H to itself (i.e., every node of H connects to both itself and every other node in H) to form a recurrent loop. Let $H_{i,t}$ denote the output of the ith node of H on recursive iteration t. If H contains 20 nodes, how many different weights of W are involved in the calculation of

$$\frac{\partial H_{i,7}}{\partial H_{k,5}} \tag{6}$$

when $i \neq k$?

$\Longrightarrow$ ..... (#21)

How many weights are involved in the same calculation when $i = k$?

$\Longrightarrow$ ..... (#22)

Clearly print your full name HERE:

2 Answers

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.
16.
17.
18.
19.
20.
21.
22.